Categorical Theories of Consciousness: Bridging Neuroscience and Fundamental Physics

Project Proposal for FQXi's *Consciousness in the Physical World*

Johannes Kleiner[1], Sean Tull[2], Quanlong Wang[2,3], and Bob Coecke[2,3]

[1]Munich Center for Mathematical Philosophy, University of Munich
[2]Topos Institute, Oxford
[3]Department of Computer Science, University of Oxford

1 Introduction

The goal of this project is to apply the powerful mathematical language of *category theory* to re-formulate and ultimately unify each of the major existing theories of consciousness. This will provide a unified perspective on these theories which is both mathematically rigorous and conceptually well-motivated. As well as this, this project aims to foster a new growing international research community on mathematical, and particularly categorical, approaches to the mind-matter relation.

In recent years, applications of category theory have led to revolutionary projects in computer science, biology, cognition, and notably physics. In the latter, the use of categories has led to a new graphical formalism for reasoning about quantum informational processes, pioneered by our PI Bob Coecke, and subsequently applied by further members of our team to yield full axiomatisations of quantum theory itself [1, 9, 32, 38]. This diagrammatic framework of *process theories*, also known as *symmetric monoidal categories*, is now being applied inside and outside of the quantum setting to a broad range of topics including natural language processing and cognition [11].

In this project, we aim to apply these 21st century mathematical tools to one of the most foundational topics across all of science: the nature of consciousness. While the topic has historically remained on the fringes of scientific discourse, over the past three decades, a growing community of researchers including not only philosophers and neuroscientists but also mathematicians, computer scientists and physicists, have begun to approach the problem, and number of major new scientific *theories of consciousness* have been developed.

One of the most promising and successful theories of consciousness so far is *Integrated Information Theory (IIT)*, developed by Giulio Tononi and collaborators [37, 33]. In recent work, members of our team have applied categorical techniques to provide an in-depth mathematical study of IIT [29, 39]. This work provided a clear new reformulation of the theory and its central algorithm which aims to capture the quality of a physical system's conscious experience, along with its quantity or 'Φ value'. While IIT has previously been only formulated for rather simple classical systems, this work allowed the theory to be extended to much more general physical settings.

Figure 1: This project's initial work on Integrated Information Theory (IIT) was featured as the cover story in the May 2nd edition of *New Scientist*, available *here* or *here*.

As well as being presented at the 2019 *Models of Consciousness* conference in Oxford, the groundbreaking nature of this project was recognised in its feature as a cover story in the May 2nd Edition of *New Scientist* [2], which explored the potential use and implications of the mathematisation of theories of consciousness, centering on our work on IIT.

Having established the viability of these categorical tools through the study of IIT, the aim of this project is now to apply them to several more of the most prominent neuroscientific theories of consciousness, namely *Predictive Processing Theory* [18, 20] and *Global Neuronal Workspace Theory* [12, 13, 30]. Additionally we will develop a categorical understanding of the framework of *Conscious Agent Networks* [19]. Specifically, our goal will be to study the essential formal structure of these theories, and then present them categorically in the language of process theories.

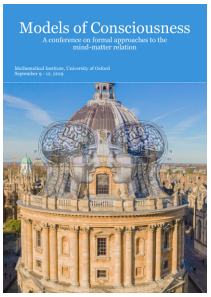
This reformulation will come with several major benefits. Firstly, thanks to its simple diagrammatic presentation, the process theory framework is easy to learn and apply, allowing new researchers from formal backgrounds to understand and study these theories quickly, fostering further research.

Secondly, any theory stated purely in process-theoretic terms becomes neutral as to the underlying physics. This allows theories previously stated in terms of classical physics, such as IIT, to be immediately applied for example to the quantum domain, as in our previous work [29, 39]. The fact that quantum mechanics is well-understood categorically [9] allows one to assess each theory's interaction with the quantum domain, and ultimately address the role that fundamental physics plays in consciousness itself.

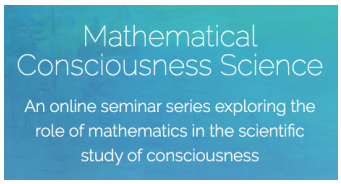
Thirdly, while each of these theories are currently presented very differently, stating them in a common categorical language will allow us to compare them much more readily, and assess what aspects of consciousness they each describe. Indeed, category theory is also a powerful language for transferring knowledge between different domains of science. A major goal of this project is to develop a precise categorical (i.e. *functorial*) description of the relation between each of these theories. Ultimately, this may allow them to be unified into a single, mathematically precise theory of consciousness.

Our team is ideally suited to this project. Bob Coecke is a world expert in the applications of process theories to both physics and AI, and Quanlong Wang and Sean Tull have both conducted DPhil and postdoctoral research within his group at the University of Oxford, which is a global centre for categorical approaches to physics. Johannes Kleiner has been heavily involved in the field of mathematical consciousness science and is now building bridges between philosophers and scientists in the context of mathematical approaches to consciousness at the Munich Center for Mathematical Philosophy. The

(a)



(b)



(c)

Figure 2: Events co-organised by members of our team. (a) *Modelling Consciousness* workshop, 2019. (b) September 2019 conference *Models of Consciousness* at Oxford. (c) Homepage for the online seminar series *Mathematical Consciousness Science*. All of these events were generously supported by FQXi grants, for which we are very grateful.

project will be based at the *Topos Institute* (*Link*), a newly formed non-profit institution which is in the process of establishing an Oxford branch right now. This Oxford branch will serve as a hub for the international applied category theory community.

Additionally, we have called for a scientific advisory committee to supervise this project, confirmed to comprise eminent researchers including Nao Tsuchiya (Neuroscience), Bechir Jarraya (Neuromodulation, GNW), Karl Friston (Predictive Processing), Gustavo Deco (Computational Neuroscience), Samson Abramsky (Computer Science) and Chetan Prakash (Conscious Agent Networks).

2 Building the Consciousness Science Community

Aside from this research goal, the second aim of this project is to foster an international community on mathematical, and particularly categorical, approaches to consciousness research, drawing on the extensive past experience of our team in such community projects.

In recent years, consciousness has become increasingly prevalent as a topic of research from those with formal backgrounds in mathematics, computer science or physics. Several members of our team have been heavily involved in the growth of this burgeoning new research field of *Mathematical Consciousness Science*, including PIs Johannes Kleiner and Sean Tull, as well as close collaborator of the group Robin Lorenz, who is currently at Oxford and will be joining the Topos Institute in September 2020.

Kleiner and Lorenz founded the online seminar series *Progress and Visions in the Scientific Study of the Mind-Matter Relation* which ran throughout 2018, hosting talks and discussions from eminent researchers from mathematics, physics and beyond. The next iteration of the online seminar series, *Mathematical Consciousness Science*, is currently taking place, being co-organised additionally by Tull. It has been highly successful with over 600 registered participants and talks attracting online audiences of over 150 live members, attended by prominent consciousness researchers including Anil Seth and Dave Chalmers.

Our team has also been involved in the co-organisation of the *Modelling Consciousness* workshops based in Dorfgastein, Austria in 2019, and online in 2020 following the COVID-19 pandemic. These involve researchers from a wide variety of backgrounds

engaging in in-depth foundational discussions on consciousness and its treatment using formal methods.

Kleiner and Lorenz were also co-organisers of the major conference *Models of Consciousness* at the Mathematical Institute of the University of Oxford in September 2019, generously funded by FQXi. This included 11 invited talks from eminent researchers within both consciousness and the mathematical sciences, including Sir Roger Penrose, and hosted close to 100 participants from all around the globe. A second installment is now scheduled for 2021 at the Center for the Explanation of Consciousness of Stanford University.

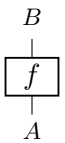
There is also currently a growing research community in *Applied Category Theory (ACT)*, within which Oxford is central, having been a host to the *ACT* conference in 2019. The *Topos Institute* (*Link*) where our project is based is one of the first dedicated centers for applied category theory, and its Oxford branch which is at present being set up will become a hub for research on the topic. As such there is a potential for a thriving new scene on applications of category theory to the study of consciousness, which our project aims to foster.

Specifically, we plan to arrange online seminars as well as a physical workshop on *Category Theory and Consciousness*, with the latter to be held in 2021. These meetings would aim to serve as the launching point for a dedicated community of researchers from applied category theory addressing theories of consciousness, and we would hope for them to become repeated events annually.

Communication with the wider community and public Along with our strong emphasis on community building comes a desire to make the results of this research project well-known to a wide audience, which we will continue to encourage, as exemplified by the *New Scientist* article [2] (Figure 1). To spread the outcomes of our seminars and workshop to a large audience we will promote them widely using our presence in the consciousness community, and by making all of their talks freely available online.

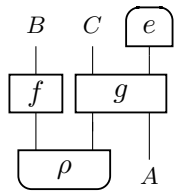
3 Process Theories

Let us now return to our core research project, and introduce the main categorical framework we will be using. A *process theory* is defined to simply be a collection of *systems*, denoted $A, B, C \dots$ and *processes* $f \colon A \to B$ between them. Such a process is typically depicted with a box as below.



In a process theory we may compose any two systems A, B to form a new system, denoted $A \otimes B$, by placing them 'side-by-side'. More generally we may compose processes together

to form so-called 'circuit diagrams' such as:



Particularly important are processes with 'no input' called *states*, such as ρ above. Process theories of an 'operational' nature typically also come with extra morphisms which allow one to *discard* any system, or prepare it in a 'maximally random' state, depicted respectively as:

Using these basic features one may go on to describe a surprisingly large number of information-theoretic concepts from both classical and quantum information theory, including *marginalisation*, *purity*, *mixedness*, *completely mixed states*, and even notions of *causality* [7]. For this reason, the diagrammatic language of such process theories is ideally suited to the study of information and causation-focused theories of science, such as IIT.

In mathematical terms, a process theory is none other than a *symmetric monoidal category* $(\mathbf{C}, \otimes)$, within which the systems are typically referred to as *objects* and the processes as *morphisms* [10]. When reasoning about processes diagrammatically we are thus employing the *graphical calculus* for monoidal categories [35].

Examples of such categories abound throughout science, being the natural formulation of any theory which describes interacting physical processes. For example, there is a category **Class** which describes classical probabilistic physics, and categories **Hilb** and **Quant** which describe pure and mixed quantum processes, respectively. States in the latter category thus correspond to quantum states in the usual sense, i.e. density matrices.

Our PI Bob Coecke pioneered the use of diagrammatic reasoning in quantum information theory, which has now developed into a large field of study known as *Categorical Quantum Mechanics* (CQM) [9], centered around the Computer Science department at Oxford, where PIs Quanlong Wang and Sean Tull have both undertaken DPhil studies and postdoctoral research.

4 Previous Work

Relation to Previous Research Our proposed project fits naturally with the prior research programs of our team, but due to its more experimental and foundational nature would be unlikely to find support within mainstream departments, without support from institutions such as FQXi.

In particular, our project continues work on applications of process theories led by Bob Coecke and (present and former) members of the Oxford quantum group Quanlong Wang and Sean Tull, as well as Johannes Kleiner's present research program on consciousness science. More specifically, it continues previous work from our team members in the articles [29, 28, 39, 3], which we describe in this section.

4.1 Integrated Information Theory

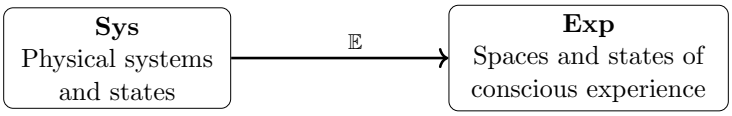
As a proof of concept of the applicability of categorical techniques to theories of consciousness, let us now describe our previous work, which formed the basis for the New Scientist Article [2].

Integrated Information Theory (IIT), developed by Giulio Tononi and collaborators, has emerged as one of the leading scientific theories of consciousness [37, 33]. At the heart of the theory is an algorithm which, based on the level of integration of the internal functional relationships of a physical system in a given state, aims to determine both the quality and quantity ('Φ value') of its conscious experience.

However, the current version of the theory ("IIT 3.0") is stated in rather long-winded terms mathematically, based on numerous examples rather than clear formal definitions. Perhaps more significantly, it can only be applied to quite simple classical physical systems. This is problematic if the theory is taken to be a fundamental theory of consciousness, and should eventually be reconciled with our present theories of physics, including quantum theory.

To resolve these issues, in the articles [29, 39] our team members Kleiner and Tull have examined the essential mathematical structure of IIT and its main algorithm, providing the latter with a crisp formalisation. Capturing the essence of the theory in this way allowed us to define a notion of generalised IIT which can be applied to very general kinds of physical systems. For example, taking quantum systems as a special case yielded the recently introduced *Quantum Integrated Information Theory* of Zanardi, Tomka and Venuti [44].

The article [29] concerns the formalisation of the IIT algorithm itself. Examining IIT 3.0 in detail, we found that any generalised IIT can be summarised as taking a class **Sys** of physical systems and specifying a mapping



(1)

into a class **Exp** consisting of mathematical structures modelling conscious experience, which we define and call *experience spaces* in [29]. This mapping sends each system S to its proposed space of possible conscious experiences $\mathbb{E}(S)$, and moreover for each state s of the system S specifies its experience $\mathbb{E}(s)$, including its intensity or Phi values $\Phi = \|\mathbb{E}(S)\|$.

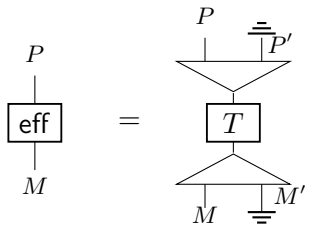
The above mapping, and thus the entirety of any generalised IIT is determined entirely by specifying the following data. Firstly, there is system class **Sys** to be considered. Next, for each system S the theory must specify an experience space $\mathbb{PE}(S)$ called its space of *proto-experiences*. The elements of $\mathbb{PE}(S)$ are ultimately combined to form full experiences of the system in the space $\mathbb{E}(S)$. Thirdly, for each system S, state s and pair of *subsystems* M, P of S, we must determine a pair of elements

$$\mathsf{caus}_s(M,P)\ ,\ \mathsf{eff}_s(M,P)\ \in\ \mathbb{PE}(S)$$

called the *cause-effect repertoire* of the system. These are intended to describe the internal dynamics of the system S. In the case of IIT 3.0, proto-experiences are given by probability distributions, while in quantum IIT they are given instead by density matrices.

Let us now see how categories enter the picture. In [39] we show that process theories provide the ideal language to describe the required features of the physical systems **Sys**, as well as the cause-effect repertoires, for both generalised IITs and for IIT 3.0 and quantum IIT specifically. Starting from any process theory **C**, the physical systems **Sys**(**C**) may be described as systems of **C** along with a *time evolution* process, as well as a set of *decompositions* of the system, which determine its subsystems as well and are used to assess how 'integrated' the overall system is.

The space of proto-experiences $\mathbb{PE}(S)$ of each such system S is essentially given by its set of states in the process theory **C**. Moreover, we show that the repertoires may be described in diagrammatic terms. For example, in the case of quantum IIT, the effect repertoire between subsystems M and P is captured by the process



where the triangles denote the decomposition of S which induce M and P respectively, and T is the time evolution. This process describes the way in which the state of M constraints that of P in the next time-step.

The repertoires for classical IIT may be recovered in similar terms. Ultimately, the article [39] provides a diagrammatic presentation of a generalised IIT which may be applied to any theory of physics specified by **C**. Taking **C** to be the category **Class** of classical physical processes yields IIT 3.0, while taking $\mathbf{C} = \mathbf{Quant}$ yields Quantum IIT instead.

There are several advantages to the reformulation of IIT in the articles [29, 39]. Firstly, the clear formalisation of the IIT algorithm allows the theory itself to be more readily understood by those from formal backgrounds, such as mathematicians, physicists and computer scientists. It also makes it clearer how the theory may be modified in future, and indeed IIT has already undergone many major changes since its initial presentation in [37]. By generalising IIT to any process theory we are now also able to more readily unify IIT with theories of fundamental physics, as exemplified by the ability of our approach to capture Quantum IIT.

Our research proposal for this project is to apply similar techniques to formalise several other major theories of consciousness in the language of process theories, in order for them to better understood, generalised and ultimately unified in future.

4.2 Further Previous Work

Further previous work that members of our team have carried out is important for the scope of this project.

First, PI Quanlong Wang along with collaborator of our team Camilo Miguel Signorelli has made use of the *ZX-calculus*, a graphical language for quantum computing [8], to model aspects of consciousness described by the philosophy of *consciousness-only* from the *Yogacara School* [43, 42] in the article [3]. This approach states that there is 'nothing but mind in the world', where mind is understood in a broad sense which in fact

captures several distinct notions of consciousness. The perception of this consciousness is modelled as a symmetric monoidal functor

$$\text{ZX-Diagrams} \to \mathbf{Hilb} \tag{2}$$

where **Hilb** is the category of finite-dimensional Hilbert spaces, describing pure quantum theory. Mental consciousnes is modelled by the process theory generated by suitable morphisms between experience spaces now described as $\mathbb{N}$-semimodules freely generated by a finite set of perceptions (impressions), either of colours, shapes, sounds, smells, tastes or touch feelings. The difference in philosophy from more physicalist approaches such as IIT, GNW, or PP is reflected in the mappings (1) and (2) going in the 'opposite direction' between the mental and physical domains. This is relevant with respect to some of the work we will carry out as part of our Subproject 3 on the *Conscious Agent Networks* model of consciousness, discussed in Section 7.

Second, PI Johannes Kleiner has investigated the implications of consciousness' specific epistemic context on the mathematical structure of formal theories of consciousness. Since consciousness is a phenomenon unlike any other studied in science, so the motivation of his work, mathematical models of consciousness might need to satisfy conditions which models of other phenomena in science might do not. In [28], he discovered that this is indeed the case, e.g., with respect to some of the *epistemic features* of conscious experience. If qualia are *ineffable*, *private* or *subjective*, a mathematical theory that seeks to address qualia needs to take into account that the epistemic access to states of the theory is fundamentally limited and introduces transformations which, similar to gauge transformations in physics, do not have observable consequences. He shows that in order for a theory to be well-defined in light of these transformations, it needs to carry a particular symmetry, whose symmetry group is given by the automorphisms of the experience spaces that constitute **Exp**, cf. Equation (1). This is particularly interesting with respect to our proposal since in the context of category theory concepts are usually defined up to unique isomorphism, and an automorphism is just an isomorphism form an object to itself. Thus some of the work in [28] might be taken as a motivation of why a categorical description of models of consciousness is *necessary* in light of consciousness unique epistemic context.

5 Subproject 1: Predictive Processing Theory

In the first part of this project, we will consider *Predictive Processing* (PP) Theory [18, 20], which aims to give a "truly unified account of perception, cognition, and action" [4], describing the brain as a complex, multi-layer prediction engine [41]. It's underlying mathematical framework is the *Free Energy Principle*, which unites Bayesian updating, risk-sensitive control and expected utility theory [21, 22, 23, 31]. It takes the form of a minimization principle which shapes and selects so-called generative models, statistical hypotheses about the hidden causes of sensory signals, and encodes the long-term average prediction error of these models. The key features of the Predictive Processing and the Free Energy Principle are depicted in Figure 3.

While originally not intended to address the problem of consciousness, the theory's unificatory ambitions have quickly led to proposals of how the theory could explain conscious experience as well [24, 26, 36]. At the heart of these proposals is the so-called *Winning Hypothesis View* [17], which states that among all of the generative models

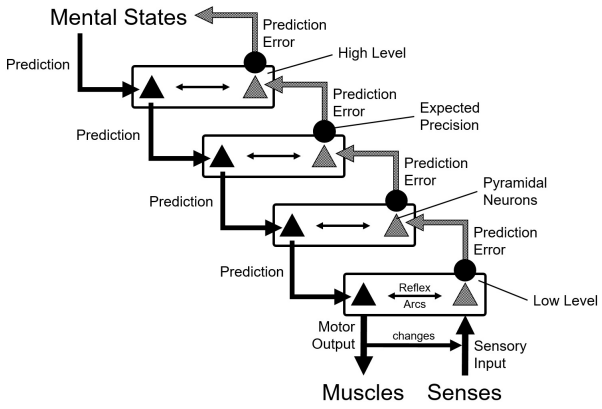


Figure 3: An illustration of key features of *Predictive Processing Theory*. The rectangles indicate hierarchically organized estimators that track features at different spatial and temporal scales, black arrows indicate top-down flow of predictions, grey arrows indicate bottom-up flow of prediction errors. The behavior all layers is determined by the mathematics of the *Free Energy Principle* [22], which involves the minimization of precision-weighted prediction error. The states of all estimators together provide a generative model which represents the predicted causes of incoming sensory signals. Picture taken from [34].

a system represents, the one which minimizes the Free Energy is also the one which determines the content of conscious experience.

In the two parts of this subproject, we aim to formulate both Predictive Processing and the Winning Hypothesis proposal in terms of process theories, providing a mathematically concise account of the formal structure of both that might inform further philosophical and scientific analysis and will provide one of three pillars for our goal of constructing a unified framework (cf. Section 8).

5.1 A Categorical Formulation of Predictive Processing

As the name of the theory suggests, Predictive Processing is concerned mainly with processes that are carried out in the brain, most notably a top-down process that propagates predictions and a bottom-up process that propagates prediction errors (Figure 3). Thus it is natural to formulate the theory in terms of Process Theories.

Since the connection to consciousness relies on specific properties of generative models, which are in turn determined by the mathematics of the Free Energy Principle, a concise mathematical description of generative models and their relations is essential. This is why, in this subproject, we aim to first define a category **Gen**, whose elements are generative models as used in Predictive Processing Theory and whose morphisms describe parthood relations between the generative models (several different models may contain the same "sub-model" that describes a certain range of conscious experiences). Similarly to the category we have used in our previous work on IIT [39], we aim to use *decompositions* in a category to capture how a generative model is represented in terms of the hierarchy of estimators explained in Figure 3.

Subsequently, we aim to define a category **Sys** which describes physical systems to which Predictive Processing Theory can be applied. Our goal is to define this category as similarly as possible to the class **Sys** we have used in our previous work on IIT, so

as to make precise comparison between the two theories possible. Finally, we define a functor

$$\mathbf{Sys} \xrightarrow{\mathbb{F}} \mathbf{Gen}$$

that explicates the Free Energy Principle. To do this, we make use of the fact that minimization principles have a natural expression in category theory in terms of what are called *limits* and *co-limits*. The functor then picks out the generative model that minimizes the Free Energy when a particular system $\mathbf{S} \in \mathbf{Sys}$ is in a particular state s.

5.2 The Winning Hypothesis as a Map between Categories

The result of Subproject 1.1. is a definition of the Predictive Processing Framework in category theoretic terms, which might already be very useful for many applications. In Subproject 1.2, we turn to Winning Hypothesis View, an extension of Predictive Processing that establishes a connection to conscious experience.

This proposed connection has become a locus of contemporary consciousness studies and is at present being subjected to thorough scrutiny. Its merits are that it allows to give convincing arguments of why and how the theory can account for apparent features of conscious experience. For example, [5] focuses on the fact that Predictive Processing Theory does not operate with exteroceptive sensory signals only, but also predicts interoceptive and proprioceptive signals. This leads to content elements of generative models (inferred latent variables in Friston's terms) whose integration with parts of the theory that describe the outside world may seem puzzling to an agent. The contributions [6] and [16], on the other hand, propose to ground an apparent mysteriousness associated with some parts of conscious experience in the fact that minimization of Free Energy leads to sparse or frugal generative models which miss a substantial amount of detail of the incoming sensory flux. Comparison of these spare representations with the rich detail of the physical world or neural structure may lead, so the proposals, to an apparent feeling of mysteriousness.

A follow-up step has recently been carried out in [17], where the winning hypothesis view is subjected to detailed scrutiny in light of the experimentally established phenomena of unconscious representation and unconscious perception (blindsight and visual form agnosia). Utilizing the more technical account of how attention is modelled in Predictive Processing, [17] proposes that a generative theory provides the apparent content of consciousness only if probed suitably by endogenous or exogenous attention, which depends on, but is not determined by, the posterior probability attributed to a theory via Free Energy. This constitutes a Dennettian multiple drafts model [15] reading of the Predictive Processing Framework.

To formalize the Winning Hypothesis View, we consider a third category **Exp** that describes (spaces of) conscious experiences. As is the case with **Sys**, we aim to define this category as close as possible to the category **Exp** used in our previous work related to IIT (Section 4). The Winning Hypothesis View can then be defined in terms of a map or functor $\mathbb{W}$,

$$\mathbf{Gen} \xrightarrow{\mathbb{W}} \mathbf{Exp}\,,$$

that specifies the content of conscious experience in terms of the content of a generative model. Depending on the specific version of the winning hypothesis view under consideration, the definition of $\mathbb{W}$ will vary, and correspondingly the properties of its image in **Exp** when conjoined with the functor $\mathbb{F}$ defined the last subproject. This will allow

for a more systematic formal analysis of the claims that are currently being made in the context of the Winning Hypothesis View and specifies the desired formulation of Predictive Processing as a model of consciousness in a Process Theory framework.

6 Subproject 2: Global Neuronal Workspace Theory

In the second part of this project, we focus on Global Neuronal Workspace Theory (GNW). This theory is, next to Integrated Information Theory (IIT) and Predictive Processing Theory (PP), the other model largely favoured by neuroscientists.

GNW is currently formulated in non-mathematical terms and stated directly in terms of brain physiology [12, 13, 30]. We are convinced that a mathematical account that allows one to formally compare the theory's predictions with IIT and PP is especially valuable here, in particular in light of Templeton's recent multi-million-dollar efforts to fund corresponding experiments [27].

We remark that a first mathematical approach to GNW has already been taken in [40], though with a different goal and perspective compared to what we propose below. We will nevertheless integrate and evaluate the results and ideas of this approach in this subproject.

Our goal in this subproject is to formulate GNW as a map or functor form some category **Sys** of systems to some category **Exp** of (spaces of) experiences, rendering the theory formally comparable to our previous work on IIT and PP.

Concerning **Sys**, we make use of the fact that similarly to classical IIT, GNW is defined relative to systems that have components, e.g. individual neurons or processing units in the brain, and whose components have connections, made up of axons, dendriters, synapses, or more general information pathways between processing units. Thus we can use similar methods as in our previous work to describe how a system $\mathbf{S} \in \mathbf{Sys}$ is comprised of components in terms of the formal language of decompositions we have introduced. This will give us a category **Sys** crafted for the application of GNW, which nevertheless is similar in structure to the **Sys** categories used for IIT and PP, respectively.

Systems so described have conscious experiences, according to the GNW theory, if two necessary conditions are satisfied. Both of these conditions refer to the topology of the graph that represents how the components of a system interact.

The first necessary condition is that the system has "two main computational spaces, each characterized by a distinct pattern of connectivity" [12, p. 56]. First, a "processing network, composed of a set of parallel, distributed and functionally specialized processors or modular subsystems subsumed by topologically distinct (...) domains with highly specific local or medium-range connections" [ibid.]. Second, a "*global neuronal workspace*, consisting of a distributed set of (...) neurons characterized by their ability to receive from and send back to homologous neurons in other (...) areas horizontal projections through long-range excitatory axons" [12, p. 56].

When translating these requirements to the process theory framework we establish, the two computational spaces can be defined in terms of different "patterns of connectivity". Concerning the latter computational space (the global neuronal workspace), we simply require that this is a network with a directed edge going into and coming out of each system. Thus in mathematical terms, the first necessary condition will look something like this:

[N1] For the system $\mathbf{S} \in \mathbf{Sys}$ to be conscious, it needs to contain two disjoint subsets N_p and N_g of components: First, a set N_p of components whose induced sub-network is a network with feed-forward connections only. Second, a set N_g (the global neuronal workspace) of components with directed edges going from this set into all components, and directed edges going to this set from all components.

To formulate this condition concisely in terms of the formal structure of **Sys** will be a first milestone of this subproject.

The second necessary condition is that "[t]he entire workspace is globally interconnected in such a way that only one such conscious representation can be active at any given time" [12, p. 58]. In mathematical terms, this can roughly be stated as follows:

[N2] The induced sub-network of N_g needs to be such that at any time t, its state 'represents' only one of the component's states.

A second milestone of our project will be to translate this condition into the process theory language we are introducing. A big part of this task will be to find a suitable definition of what constitutes a representation in [N2] in category theoretic terms.

If both necessary conditions [N1] and [N2] are satisfied at a particular time t, the GNW model claims that the system $\mathbf{S}$ is conscious of the "perceived object, event, narrative or scenario" represented in the global workspace network N_g. Due to the directed edges from N_g to the components, the state of one of the components may be made "directly available in its original format to all other workspace processes" [14, p. 15]. By combining the necessary conditions [N1] and [N2], we can thus define a map or functor

$$\mathbf{Sys} \xrightarrow{\mathbb{G}} \mathbf{Exp}\,,$$

which specifies the conscious experience of a system $\mathbf{S}$ in a particular state according to GNW.

We remark that both [N1] and [N2] will be refined in the course of this project. On the one hand, we will discuss and evaluate our formalization ideas with the GNW community (qua one of our many community-building projects and our advisory board, cf. Section 2), in order to ensure that the formal statements on which we base our mathematical constructions express exactly what the model intends to express. On the other hand, our choices will also be informed by considerations of which definitions would be *natural* in a process theoretic description of the systems that GNW is typically applied to (viz. parts of the brain).

7 Subproject 3: Conscious Agent Networks

In the final subproject of this paper, we focus on the Conscious Agent Network Model introduced in [25]. Unlike IIT and PP, this model is based on idealistic metaphysics.

The core ingredient in this model is a mathematical description of what constitutes a conscious agent (Figure 4b), and how conscious agents can interact via processes of action and perception. While originally described in terms of measureable spaces and Markov kernels, we think that this model is ideally suited to be defined in terms of a category **Can** whose objects are experience and action spaces (the former defined similarly to our category **Exp** introduced before), and whose morphisms are the action and perception processes.

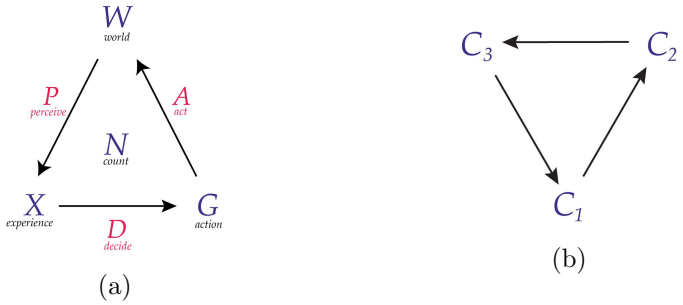


Figure 4: (a) A conscious agent as modelled in the *Conscious Agent Network* theory. W, X and G describe spaces that represent the external world, experience and action. P, A and D describe processes of perception, action and decision. (b) Conscious agents can interact using the former two processes. To every individual agent, the interaction looks as if there is an external world, though there in fact are only conscious agents. Picture from [25].

Combining this with the results of the previous project, we hope to make progress in understanding how the interaction of conscious agents as described by $\mathbb{W}$ can give rise, for every individual agent, to the appearance of an external world. Building on the previous work in [25], we hope to explicate this idea in terms of a mapping

$$\mathbf{Can} \xrightarrow{\mathbb{W}} \mathbf{Sys}\,,$$

where as before **Sys** represents physical systems. For this goal we will utilise the fact that both **Can** and **Sys** will be defined in terms of process theories.

8 Major Goal: Unifying Theories of Consciousness

Overall, the motivation for this project is to help to unify each of these theories of consciousness. Leaving aside Conscious Agent Networks for now, we hope to achieve this by defining each of the neuroscientific theories IIT, PP and GNW in terms of functors of the form

$$\mathbf{Sys} \longrightarrow \mathbf{Exp}\,,$$

where **Sys** and **Exp** are process theories (i.e. categories) that describe physical systems and conscious experience, respectively.

As the mathematical structure of the theories we consider is quite thorough, we expect that the categories **Sys** and **Exp** will differ in definition when adapted to each of them. A goal of our project is to investigate in how far these differing details can be removed without changing the theories under consideration substantially.

Concerning **Sys**, we expect differences to originate from the different emphases the theories put on the features of brain physiology they consider. For example, while IIT requires features determining how integrated a system's information processing is, GNW is more concerned with how 'global' it is. We expect that it is possible to define one large system category **Sys** on which all of these theories can operate.

Concerning **Exp**, we expect the differences to have a more fundamental nature, originating from the fact that the various theories aim to address *different conceptions of consciousness*. For example, GNW is built, primarily, on experimental insights regarding whether a subject perceives a stimulus consciously or not. In contrast, IIT is originally

built on the 'level of wakefulness' conception of conscious experience and in its most recent formulations aims to describe the whole conscious experience of a system all at once. Finally, Predictive Processing Theory is concerned mainly with cognitive features of perception (e.g. the relation of action, perception and attention), with the link to conscious experience only having been added subsequently and aiming only at the content of conscious experience.

While it is very obvious that these various conceptions of consciousness are not completely independent, it is far from obvious whether they actually aim to describe the same parts or features of conscious experience. In fact, one could take the point of view that the various theories describe different complementary aspects of conscious experience and may hence point at one underlying hidden theory of consciousness.

The last goal of this proposal aims to investigate this possibility, both conceptually and mathematically. To this end, we aim to define a category **Exp** that unites the various **Exp**s that have been defined in the previous subprojects both mathematically and formally. The goal is to define a mathematical representation of experience which incorporates and explicates the distinctions that underlie the various target phenomena of the theories we have considered previously, containing limiting or restriction procedures that allow to go from the fully category to any of the specific categories that are being used in the above theories.

We believe that positive results in direction would have substantial benefits for the scientific study of consciousness and might lead the path to a coherent and comprehensive theory of consciousness, integrating many of the ideas that have so far developed in competition.

References

[1] Samson Abramsky and Bob Coecke. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425. IEEE, 2004.

[2] Michael Brooks. Is the universe conscious? It seems impossible until you do the maths. *New Scientist*, May 2nd, 2020. Available at https://www.newscientist.com/article/mg24632800-900-is-the-universe-conscious-it-seems-impossible-until-you-do-the-maths/.

[3] Quanlong Wang Camilo Miguel Signorelli. A compositional model of consciousness based on subjectivity as a fundamental feature of nature. *Pre-print.*, 2020.

[4] Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press, 2015.

[5] Andy Clark. Consciousness as generative entanglement. *The Journal of Philosophy*, 116(12):645–662, 2019.

[6] Andy Clark, Karl Friston, and Sam Wilkinson. Bayesing qualia: consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9-10):19–33, 2019.

[7] Bob Coecke. Terminality implies no-signalling... and much more than that. *New Generation Computing*, 34(1-2):69–85, 2016.

[8] Bob Coecke and Ross Duncan. Interacting quantum observables: Categorical algebra and diagrammatics. *New Journal of Physics*, 13, 2011.

[9] Bob Coecke and Aleks Kissinger. *Picturing quantum processes*. Cambridge University Press, 2017.

[10] Bob Coecke and Eric Oliver Paquette. Categories for the practising physicist. In *New structures for physics*, pages 173–286. Springer, 2010.

[11] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

[12] Stanislas Dehaene, Jean-Pierre Changeux, and Lionel Naccache. *The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications*, pages 55–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[13] Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24):14529–14534, 1998.

[14] Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of conscious-

ness: basic evidence and a workspace framework. *Cognition*, 79(1):1 – 37, 2001. The Cognitive Neuroscience of Consciousness.

[15] DC Dennett. Consciousness explained (p. weiner, illustrator). *New York, NY, US: Little, Brown and Co*, 1991.

[16] Joe E Dewhurst and Krzysztof Dolkega. Attending to the illusion of consciousness. *Forthcoming*, 2020.

[17] Krzysztof Dolkega and Joe E Dewhurst. Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, pages 1–26, 2020.

[18] Harriet Feldman and Karl Friston. Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215, 2010.

[19] Chris Fields, Donald D Hoffman, Chetan Prakash, and Manish Singh. Conscious agent networks: Formal analysis and application to cognition. *Cognitive Systems Research*, 47:186–213, 2018.

[20] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

[21] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

[22] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[23] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, 2006.

[24] J Allan Hobson and Karl J Friston. Consciousness, dreams, and inference: The cartesian theatre revisited. *Journal of Consciousness Studies*, 21(1-2):6–32, 2014.

[25] Donald D. Hoffman and Chetan Prakash. Objects of consciousness. *Frontiers in Psychology*, 5:577, 2014.

[26] Jakob Hohwy. Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, 3:96, 2012.

[27] Philip Ball. https://www.mindcoolness.com/blog/bayesian-brain-predictive-processing/. Neuroscience readies for a showdown over consciousness ideas, 2019.

[28] Johannes Kleiner. Mathematical models of consciousness. 2019. Forthcoming in Entropy.

[29] Johannes Kleiner and Sean Tull. The mathematical structure of integrated information theory. *arXiv preprint arXiv:2002.07655*, 2020.

[30] George A Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

[31] Thomas Metzinger and Wanja Wiese. *Philosophy and Predictive Processing*. Number 978-3-95857-138-9. MIND Group, 2017.

[32] Kang Feng Ng and Quanlong Wang. A universal completion of the zx-calculus. *arXiv preprint arXiv:1706.09877*, 2017.

[33] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol*, 10(5):e1003588, 2014.

[34] Post on the website https://www.mindcoolness.com/blog/bayesian-brain-predictive-processing/. The bayesian brain: An introduction to predictive processing, 2018.

[35] Peter Selinger. A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer, 2010.

[36] Anil K Seth and Hugo D Critchley. Extending predictive processing to the body: emotion as interoceptive inference. *Behav. Brain Sci*, 36(3):227–228, 2013.

[37] Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5(1):42, 2004.

[38] Sean Tull. A Categorical Reconstruction of Quantum Theory. *Logical Methods in Computer Science*, Volume 16, Issue 1, January 2020.

[39] Sean Tull and Johannes Kleiner. Integrated information in process theories. *arXiv preprint arXiv:2002.07654*, 2020.

[40] Rodrick Wallace. *Consciousness: A Mathematical Treatment of the Global Neuronal Workspace Model*. Springer, 2005.

[41] Wanja Wiese and Thomas Metzinger. Vanilla PP for philosophers: A primer on predictive processing. 2017.

[42] Xuanzang, Francis H.Cok and Vasubandhu. *Three Texts on Consciousness Only*. Numata Center for Buddhist Translation and Research, Berkeley, 1999.

[43] Xuanzang, Tat Wei and Vasubandhu. *Cheng Wei Shi Lun; The Doctrine of Mere-Consciousness*. Ch'eng Wei-shih Lun Publication Committee, Hong Kong, 1973.

[44] Paolo Zanardi, Michael Tomka, and Lorenzo Campos Venuti. Towards quantum integrated information theory. *arXiv preprint arXiv:1806.01421*, 2018.